Class Activation Map Analysis of Fine-tuned Deep Networks for Pokemon Classification

Project Proposal

Shivani Shrikant Naik

Pulkit Saharan

Problem Definition

For many years, Convolutional Neural Networks were treated as black boxes with inputs and outputs, but we didn't know what the network was learning. Class Activation Map (CAM) aids in understanding what features the network is learning and which areas are prioritized in order to distinguish classes from one another.
We intend to use CAMs in a pokemon classification problem setting to learn more about the internal workings of a deep CNN. This problem statement will allow us to fine-tune pre-trained networks on a new classification task, followed by an analysis of the newly trained network to see what features it is learning.

Dataset

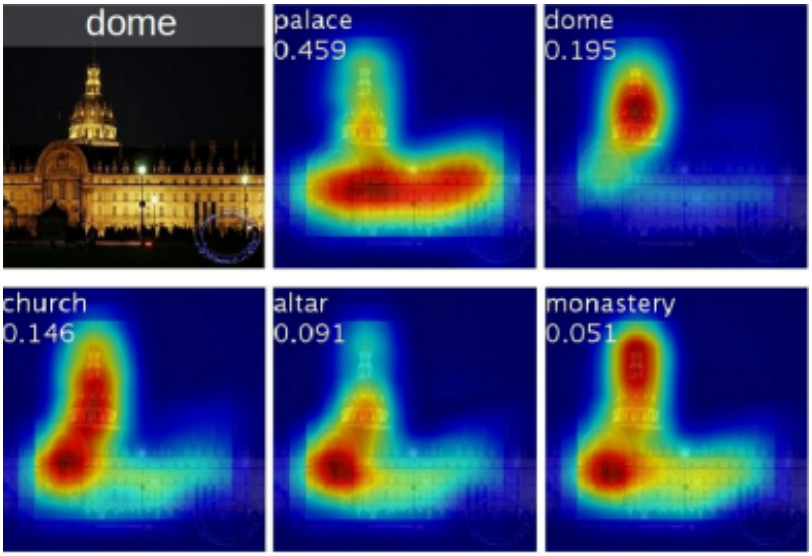
Dataset Link: https://www.kaggle.com/datasets/thedagger/pokemon-generation-one

The dataset consists of 151 pokemons, with 60 images each per pokemon. Out of these 151 pokemons, we plan to select 10 classes for our analysis, with some distinctive and some similar looking classes.

Class Activation Maps

Class Activation Maps are a powerful and intriguing technique in the field of Neural Network classification. This technique provides more insight into the network's learning and aids in debugging. In addition to predicting the class using the network, CAMs show us which part of the image the network is most interested in. Activation Maps show the major regions of an image for a specific prediction. The user receives an object localization of the predicted class using this technique without having to explicitly label the bounding box for this object.
Given below are a few examples to show how CAMs use the salient region of an image in prediction.

Activation Map for Class Siberian Husky (ImageNet Class #250)



Proposed Work

We intend to experiment and use transfer learning to fine-tune a pre-trained neural network such as VGG-16 or Resnet, resulting in a model that can classify various pokemons. A pre-trained network is less likely to overfit on small datasets and is also faster to train.

This network will then be used to generate Class Activation Maps for various classes. If time allows, we will also attempt to generate Grad-CAMs for this network.

References

1. Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2019, December 3). *Grad-cam: Visual explanations from deep networks via gradient-based localization*. arXiv.org. Retrieved November 30, 2022, from https://arxiv.org/abs/1610.02391

2. Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., & Torralba, A. (2015, December 14). *Learning deep features for discriminative localization*. arXiv.org. Retrieved November 30, 2022, from https://arxiv.org/abs/1512.04150